

**Dimemas overview**

Jesús Labarta, Judit Gimenez  
Jordi Caubet, Francesc Escale  
CEPBA-UPC

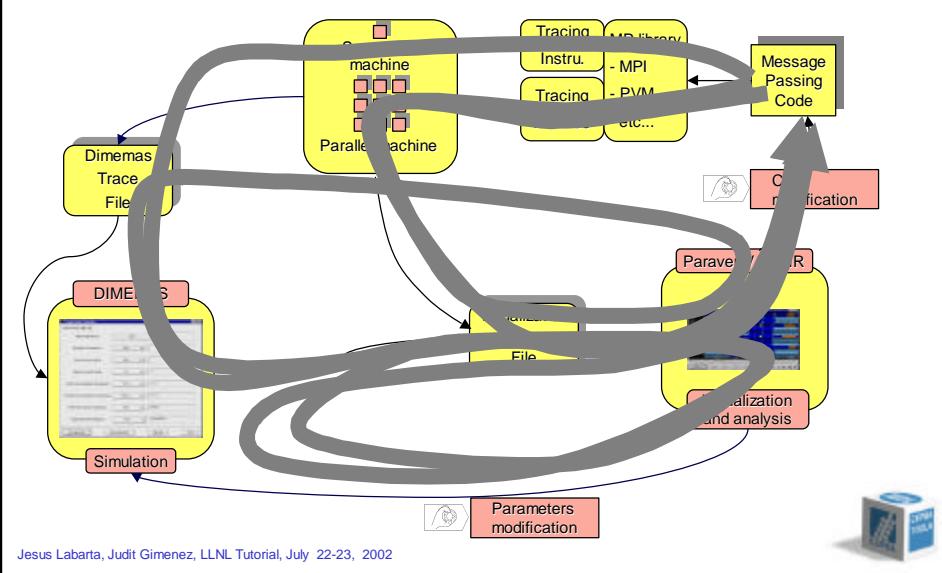
Technology Transfer      Research      Training      Mobility of Researchers  
User Support      Education      HPC Facilities      Parallel Expertise

## Index

- The model
- Qualitative validation
- Quantitative Validation
- Examples
- New developments

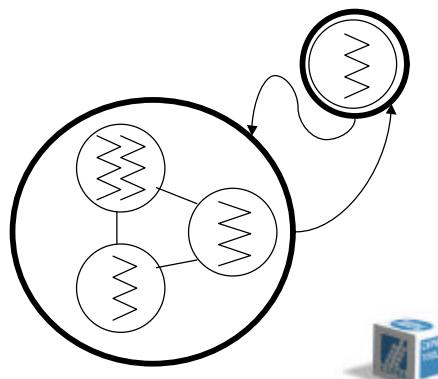
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Dimemas



## Tracefile

- Characterises application
  - Sequence of resource demands for each task
  - Sequence of events: communication
- Application model

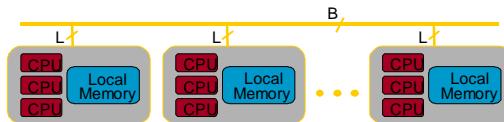


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

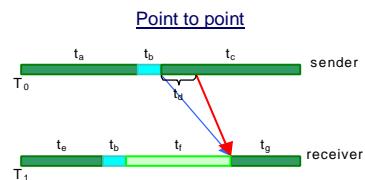
## Simulated Architecture

### ■ “Abstract” architecture

- Simple/general
- Fast simulation
- Key factors influencing performance
  - ✓ Local/remote Latency/BW
  - ✓ Injection mechanism
    - #links
    - half/full duplex
  - ✓ Bisection BW, contention
  - ✓ Basic MPI protocols



### ■ Multiprogrammed workload



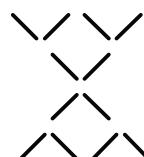
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



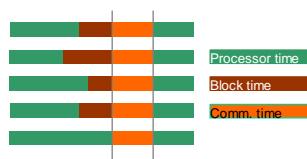
## Simulated Architecture

### ■ Collective communication model

- Fan-in / Fan-out
- Size of message
- Lin / log / const
- Barrier



Collective



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Collective Communication Model

### ■ Communication time

$$\text{Time} = \left( \text{Latency} + \frac{\text{Size}}{\text{Bandwidth}} \right) * \text{MODEL\_FACTOR}$$

### ■ Model factor

Model	Factor
Null	0
Constant	1
Linear	P
Logarithmic	$N_{\text{steps}} = \sum_{i=1}^{\lceil \log_2 P \rceil} \text{steps}_i, \text{steps}_i = \left\lceil \frac{C}{B} \right\rceil$

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Instrumentation

### ■ Dimemas instrumentation

- MPIDtrace
  - ✓ Run the same was as OMPItrace

### ■ Tracefile generated by Paraver

- The Paraver trace should have been obtained with dedicated resources

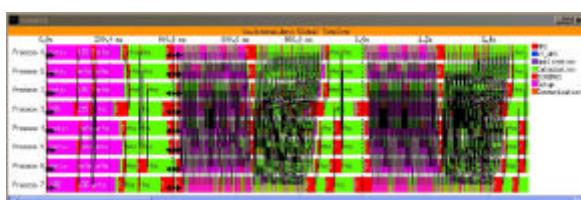


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Qualitative validation



Dedicated machine



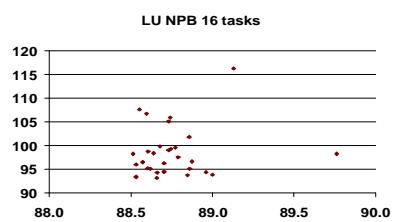
Shared  
environment  
+ Dimemas

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

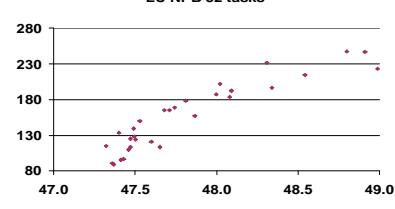


## Stability validation

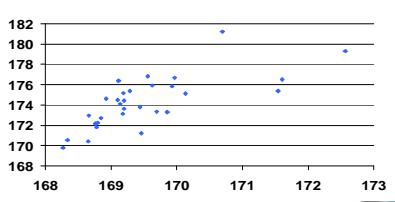
- On loaded system, 30 times
  - trace & measure elapsed time
  - predict time for dedicated system



LU NPB 32 tasks



LU NPB 8 tasks

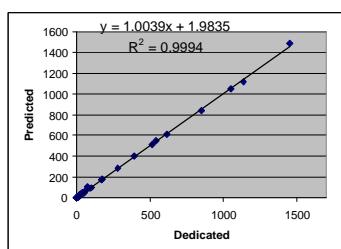
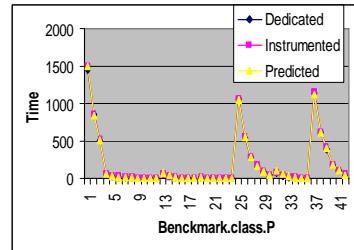


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## NAS Benchmarks

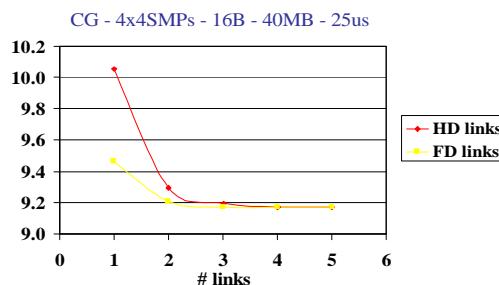
- BT, CF, FFT, MG, IS LU, SP
- Class W and A
- P = 8.32
- Target machine model
  - L = 27, BW = 80, B =  $\infty$



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Bandwidth Trade-offs

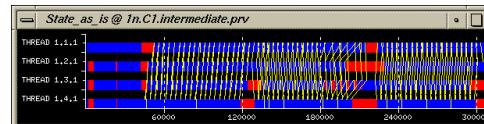
- Injection mechanism



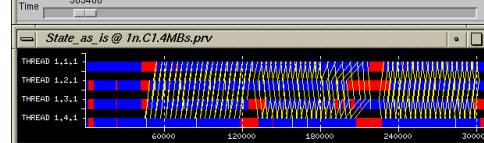
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Bandwidth Trade-offs

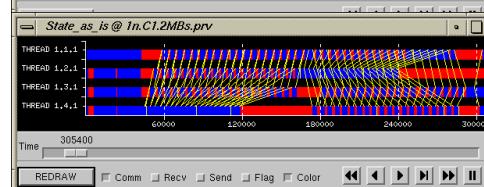
- L=50  $\mu$ s , BW= 8 MB/s



- L=50  $\mu$ s , BW= 4 MB/s



- L=50  $\mu$ s , BW= 2 MB/s



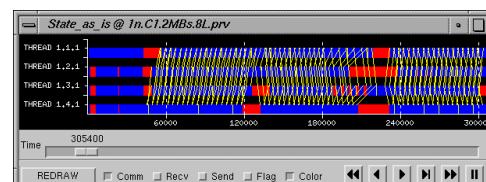
====> 2 MB/s is a problem ...



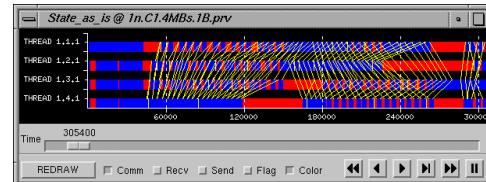
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Bandwidth Trade-offs

- L=50  $\mu$ s , BW= 2 MB/s but  
8 links to network per  
node



- L=50  $\mu$ s , BW= 4 MB/s but  
1 bus

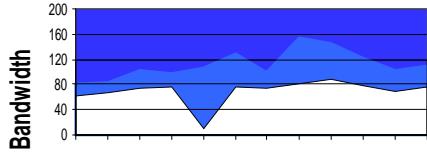


... trade off raw bandwidth -  
bisection/connection BW

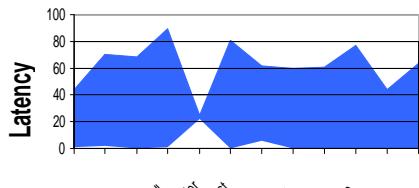


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## System characterization



< 10% error regions



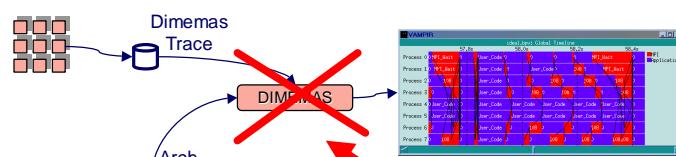
1 half duplex link !!!

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

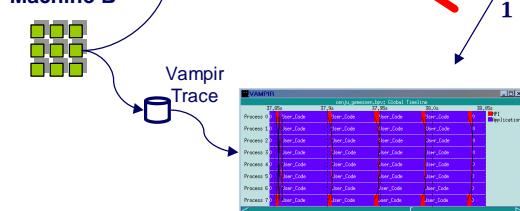


## Understanding architectures

Machine A



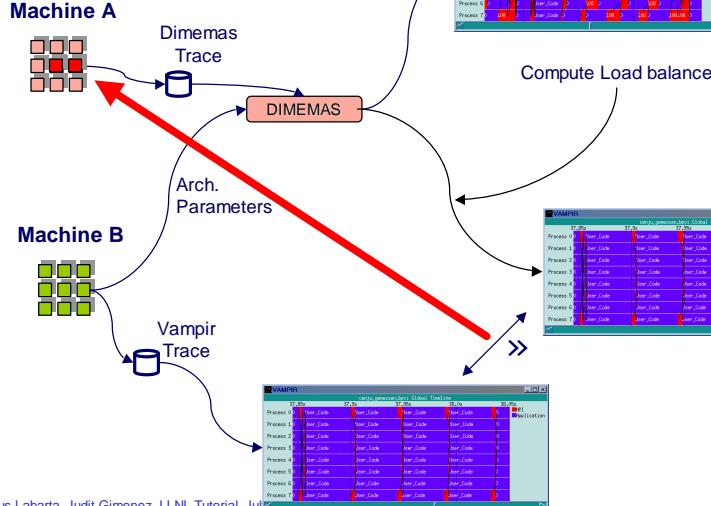
Machine B



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



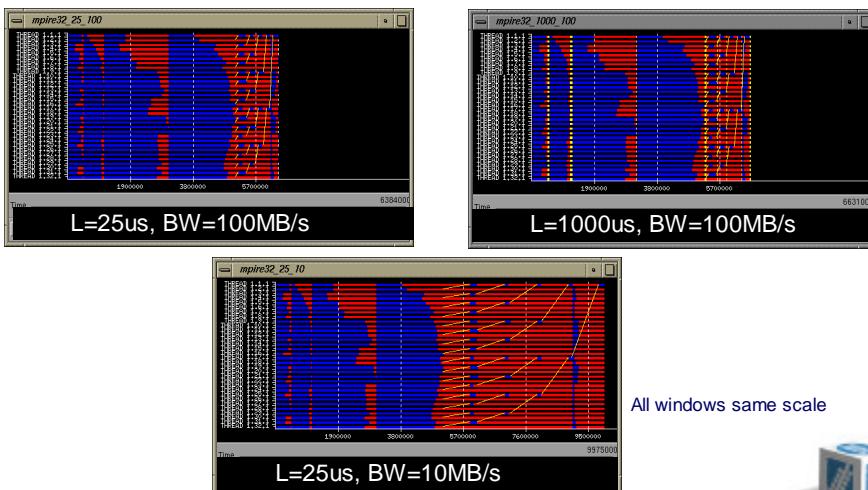
## Understanding architectures



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Understanding applications (MPIRE)

- 32 procs (no network contention)

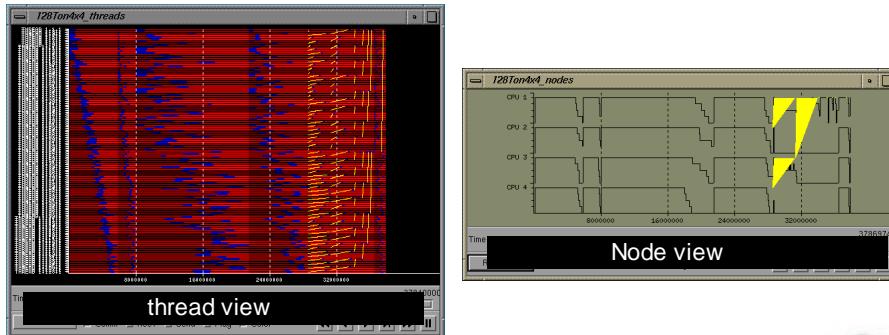


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Understanding applications (MPIRE)

### ■ Cluster of SMPs

- 4nodesx4, 1 link
- 128 p, L=25, BW=100, no network contention



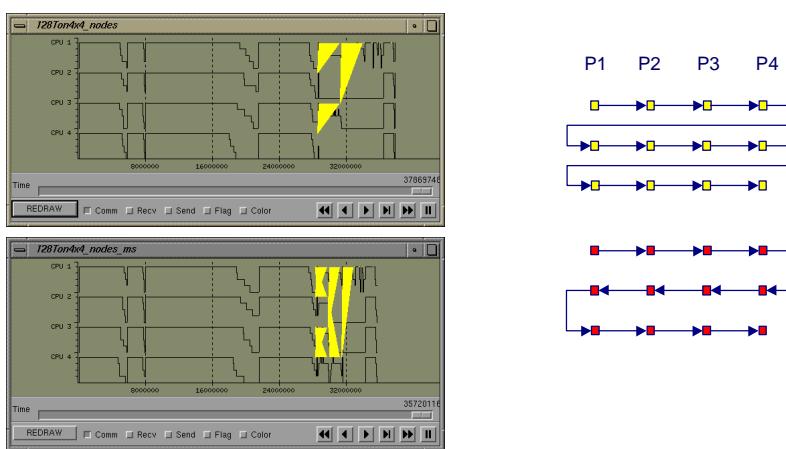
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Understanding applications (MPIRE)

### ■ Mapping influence

- 128 p, L=25, BW=100, 4nodesx4, 1 link, no network contention



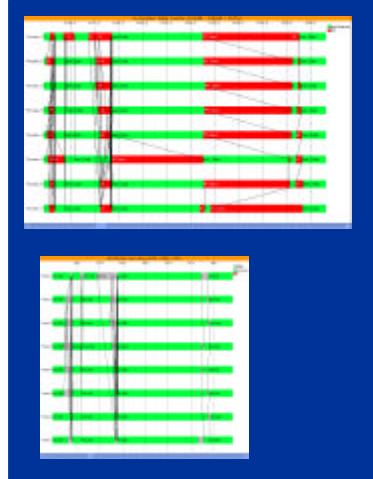
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Understanding architectures

### ■ Unexpected behavior

- IFS on E10000 (MPICH)\*
- IBM SP (@KTH)



\* Courtesy FECIT

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Sensitivity analysis

### ■ How sensitive is my program to bandwidth, latency,injection mechanism, routine optimization?

- How much do they influence the execution time?
  - ✓ ST-ORM: "GRID" tool for specifying studies (montecarlo, optimization,parametric studies...), generating and submitting the jobs, collecting and analyzing results.
- Is it possible to build simple models of program performance?

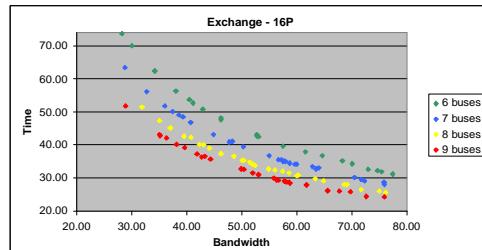
### ■ Which parts of my program are more sensitive to bandwidth, latency,routine optimization?

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Montecarlo studies

### ■ Contention

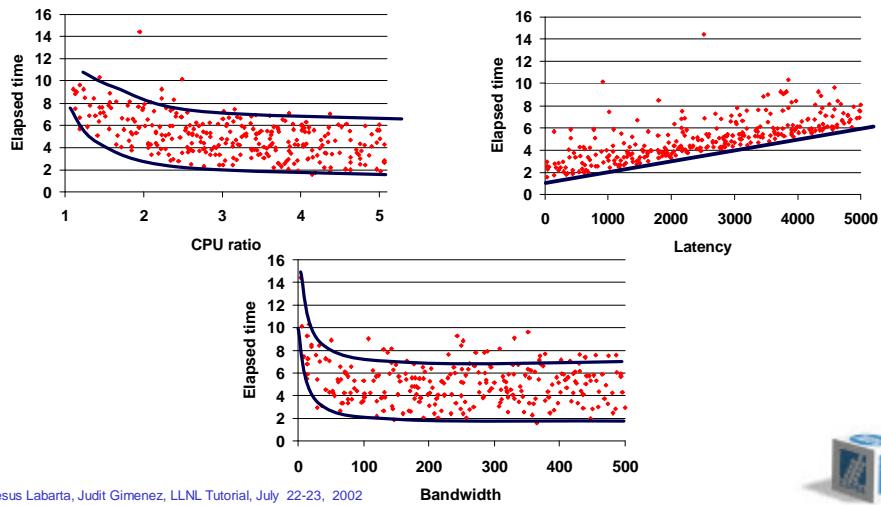


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Montecarlo studies

### ■ MPIRE: 32 CPUs (no network contention)

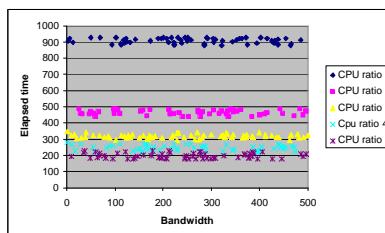
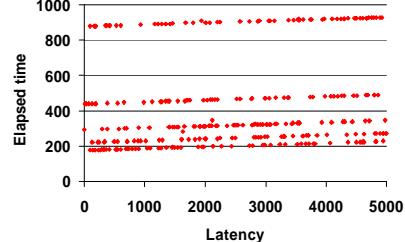
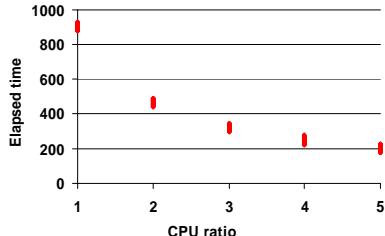


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Montecarlo studies

### ■ SCF: 32 CPUs (no network contention)

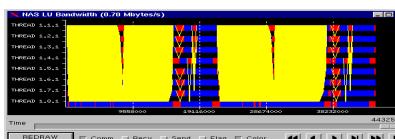


Essentially Sensitive to raw processor performance

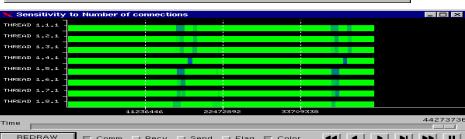


Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

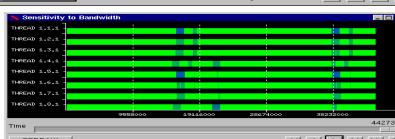
## Program section sensitivity



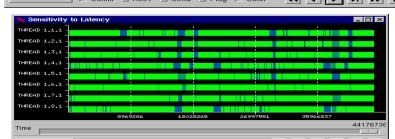
to contention



to bandwidth



to latency



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002

## Metasim + Dimemas

### ■ Convolution

- Observed performance is a convolution of
  - ✓ algorithm characteristics
    - Instructions
    - Communication demands
  - ✓ machine characteristics
    - Processor
    - Communication

### ■ Tracing is an attempt to deconvolve the algorithm and machine characteristics from an actual run to later convolve (Metasim/Dimemas) with the hypothetical target machine characteristics

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Metasim + Dimemas

### ■ Cooperation with Allan Snavely (SDSC)

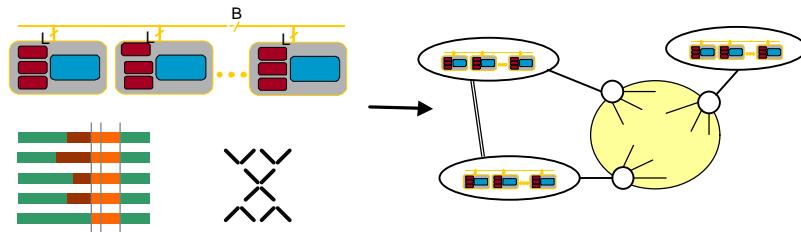
### ■ Estimate CPU ratios between tracing and target machine

- Based on simple instruction level simulation
- Started by simple hypothesis
  - ✓  $IPC \propto \text{bandwidth}$

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Dimemas GRID: model extension

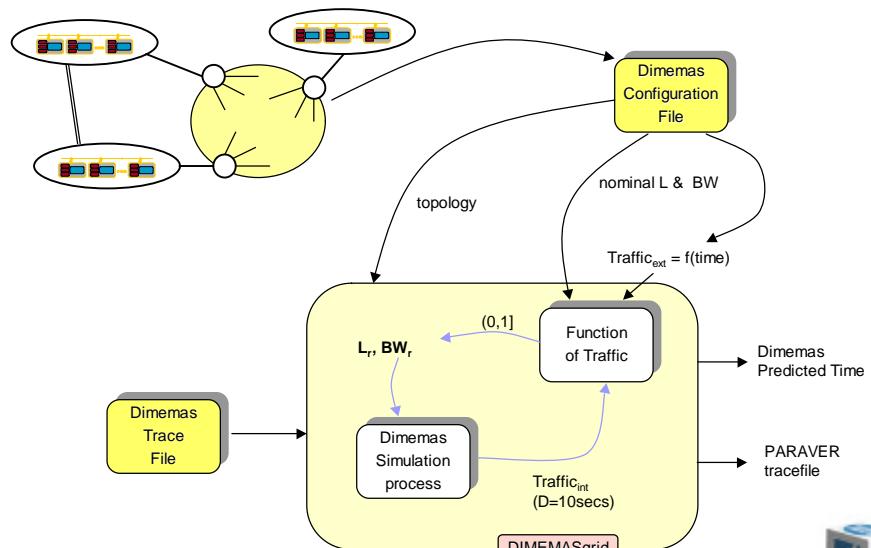


- Dedicated connections
- External network
  - Variation on effective bandwidth due to traffic
- Collective communication extension

Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Dimemas GRID



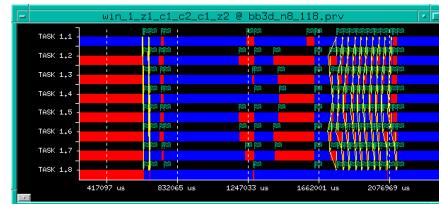
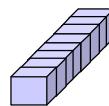
Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002



## Dimemas GRID: example

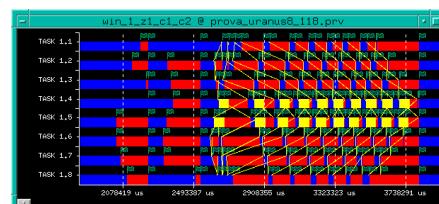
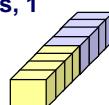
- URANUS

- 1 machine, 8 nodes, 1 proc/node



- latency=30 msecs, BW=0.8 MB/S

- 2 machines, 4 nodes, 1 processor/node



Jesus Labarta, Judit Gimenez, LLNL Tutorial, July 22-23, 2002